Abstract

Non-parametric methods like nearest neighbor (NN) classification and Parzen-window based density estimation are more general than parametric methods because they do not make any assumptions regarding the hypothesis class or probability distribution form Further, they show good performance in practice with enough number of samples Curse of dimensionality and high computational requirements are the two major problems in applying non-parametric techniques Because of curse of dimensionality, many classifiers become severely biased with a small training set This problem is prominent with non-parametric methods where the demand for a larger number of samples grows exponentially with the dimensionality of the feature space It is widely believed that the size of the training set needed to achieve a given classification accuracy would be prohibitively large when the dimensionality of data is high There exist several remedies like feature selection and bootstrapping to solve the curse of dimensionality problem Computational requirements can be reduced by prototype selection, building an index over the training set, etc

The thesis offers a novel solution for the problems listed above especially for NN classifiers which are based on (i) increasing the training set size by adding artificial patterns to it, (ii) finding a compact representation for the entire synthetic set (consisting of both the given and artificial patterns), and (iii) finding efficient ways of doing NN classification which directly work with the compact representations Since the synthetic training set used is large, this can reduce the curse of dimensionality problem Because of compact representations and efficient NN algorithms, computational requirements are reduced

The process of generating artificial patterns is called pattern synthesis Pattern Synthesis can be done broadly in two ways *viz , model based pattern synthesis* and *instance based pattern synthesis*

Model based pattern synthesis first derives a model based on the training set and uses this to generate patterns The model derived can be a probability distribution or an explicit mathematical model like a hidden Markov model This method can be used to generate as many patterns as needed There are two drawbacks of this method First, any model depends on the underlying assumptions and hence the synthetic patterns generated can be erroneous Second, it might be computationally expensive to derive the model Another argument against this method is that for many pattern recognition tasks the model itself can be used without generating any patterns at all

Instance based pattern synthesis on the other hand, uses the given training patterns and some of the properties of the data It can generate only a finite number of new patterns Computationally this can be less expensive than deriving a model This is especially useful for NN Classifiers which directly use the training instances

This thesis mainly deals with instance based pattern synthesis Two broad instance based synthesis methods, *viz , partition based pattern synthesis* and *overlap based pattern synthesis* are presented Since the number of synthetic patterns that can be generated can be exponential in the number of given original patterns, explicit generation of synthetic patterns is not feasible Two compact representations *viz ,* partitioned pattern count tree (PPC-tree) and overlap pattern graph (OLP-graph) to store the entire synthetic sets for partition based pattern synthesis and overlap based pattern synthesis, respectively are presented

Efficient NN classification techniques based on divide-and-conquer strategies and dynamic programming techniques, which use the compact representations directly are presented Their superiority over conventional and established methods are experimentally demonstrated over various standard datasets Also, relationships among various NN based methods using synthetic patterns are established

An $O(1)$ NN classifier with synthetic patterns which is based on approximate NN

search is presented Its superiority also is established experimentally

Finally, partition based synthetic patterns are used with Parzen-window based density estimation for network intrusion detection The method is based on novelty detection approach and directly works with the corresponding compact representations Experimentally, this method is shown to perform well

The methods presented are suitable for large and high dimensional datasets like those in data mining applications